RESEARCH PAPER

A Knowledge Graph Based Approach to Social Science Surveys

Jeff Z. Pan[1,2†], Elspeth Edelstein[3], Patrik Bansky[2] & Adam Wyner[4]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9YL, UK

[2]Department of Computing Science, University of Aberdeen, Aberdeen AB24 3FX, UK

[3]School of Language, Literature, Music and Visual Culture, University of Aberdeen, Aberdeen AB24 3FX, UK

[4]School of Law and Department of Computer Science, Swansea University, Swansea, West Glamorgan SA2 8PP, UK

Keywords: Intelligent survey system; Dynamic and informative system; Knowledge graph; Linguistic grammaticality judgements

Citation: Pan, J.Z., et al.: A knowledge graph based approach to social science surveys. Data Intelligence 3(4), 477-506 (2021). doi: 10.1162/dint_a_00107

Received: January 10, 2021; Revised: July 1, 2021; Accepted: July 6, 2021

ABSTRACT

Recent success of knowledge graphs has spurred interest in applying them in open science, such as on intelligent survey systems for scientists. However, efforts to understand the quality of candidate survey questions provided by these methods have been limited. Indeed, existing methods do not consider the type of on-the-fly content planning that is possible for face-to-face surveys and hence do not guarantee that selection of subsequent questions is based on response to previous questions in a survey. To address this limitation, we propose a dynamic and informative solution for an intelligent survey system that is based on knowledge graphs. To illustrate our proposal, we look into social science surveys, focusing on ordering the questions of a questionnaire component by their level of acceptance, along with conditional triggers that further customise participants' experience. Our main findings are: (i) evaluation of the proposed approach shows that the dynamic component can be beneficial in terms of lowering the number of questions asked per variable, thus allowing more informative data to be collected in a survey of equivalent length; and (ii) a primary advantage of the proposed approach is that it enables grouping of participants according to their responses, so that participants are not only served appropriate follow-up questions, but their responses to these questions may be analysed in the context of some initial categorisation. We believe that the proposed approach can easily be applied to other social science surveys based on grouping definitions in their contexts. The knowledge-graph-based intelligent survey approach proposed in our work allows online questionnaires to approach face-to-face interaction in their level of informativity and responsiveness, as well as duplicating certain advantages of interview-based data collection.

† Corresponding author: Jeff Z. Pan (Email: j.z.pan@ed.ac.uk; ORCID: 0000-0002-9779-2088).

1. INTRODUCTION

With an increasing variety of advanced artificial intelligent techniques being developed, researchers are starting to investigate how to apply artificial intelligence techniques in the social sciences. This paper is about using knowledge graph to build a dynamic, intelligent system for social science surveys by way of a particular instantiation of a survey about linguistic grammaticality judgements.

The paper is set within the context of surveys in the social sciences, so we briefly consider some general, relevant points selected from the much wider domain of research on survey methodology [1, 2, 3, 4, 5]. This review is followed by a more specific discussion tied to our proposal. For our purposes, surveys may be broadly divided into two broad *survey modes*—questionnaires (group, mail, drop-off, and online), and interviews (personal, phone, and online). Questionnaires are typically structured, asking a series of specific, statically ordered questions to which the participant may respond with provided alternative answers, e.g., *multiple choice* questions; questionnaires may be conducted by paper and pencil or online. Interviews are typically less structured, more dynamic, exploratory, interactive, and in person. The former is less resource intensive than the latter, but the latter may be more informative, depending on researchers' purposes, materials, and analysis.

The distinction between surveys and interviews need not be hard and fast. For instance, as is relevant to our approach, an online questionnaire may ask specific questions with provided alternative answers, yet be dynamic, interactive, and personalised in that subsequent questions are conditional on previous answers for that particular participant. For surveys in social science, issues about population and access to the population are relevant, bearing on sample size, literacy, language, cooperation, geography, and (for online surveys) Internet usage. In addition, construction of a survey must take into account the capability of participants to answer the questions: Are they likely to be sufficiently insightful to informatively complete the survey? In the current work, the topic of investigation helps to address these issues.

Turning to composition, the survey instrument requires careful planning about the content, wording, format, and order of questions. For instance, the order of questions may be without alternatives, where all questions must be answered and in order; alternatively, *skip patterns* may be deployed, where a skip pattern is a question that is given dependent on the response to a previous question while skipping over other questions. In both instances, a decision tree is used and most commonly allied with a *database* to store the underlying data. As will be shown, our approach can in part be taken to be an advanced, articulated form of skip pattern questioning. However, in addition to utilising a decision tree, we take the novel approach of integrating it with a *knowledge graph*, which represents domain knowledge and enables a high degree of flexibility, fine-grainedness, complexity, and extensibility. Moreover, the data must be analysed in such a way as to address the goals of the research, and the survey must be constructed so as to serve the analysis and goals. For example, if one is seeking correlations between income and psychological impacts, one samples a relevant population, and then abstracts over individuals in the analysis. In contrast, if one wishes some further analysis of individual variation, as is done in this study, then the structure of the survey and the record of the data must serve such ends.

Finally, the underlying technology may be relevant. It is common to serve a survey by way of a program online that provides questions and responses which are extracted from a database. In the novel approach developed here, rather than a database, a knowledge graph of fine-grained, domain-specific information is used and remixed, which has advantages in terms of flexibility and complexity of questions as well as depth of analysis. In the following sections, we further develop these observations, selecting and developing a specific approach to surveys in social science which serves the goals of the particular research topic.

Our previous work [6] presents the architecture of a knowledge driven intelligent survey system, in which questions are ordered not only to pique participants' interest, but also to optimise data collection in order to test particular hypotheses. The idea is to use a knowledge graph not only as a semantic bridge between humans and computational systems, but also to facilitate customisability, transmission, re-usability, explainability, extensibility, and reasoning.

The system architecture has three different components. These components are exposed to their relevant users:

(1) survey participants;
(2) domain experts; and
(3) knowledge engineers.

The system is instantiated for linguistic surveys about grammaticality judgements. The *survey participants* simply answer the questions: Their role is to judge whether sentences are acceptable or unacceptable. The *domain experts*, such as linguists, customise the knowledge structure to fit their needs. The *knowledge engineers* construct the basis of the semantic structure, which the linguist customises and instantiates. For example, a knowledge driven intelligent survey system for linguistics provides information about syntactic relationships and linguistic features for each sentence. From participant submissions, researchers are able to see detailed information about these syntactic relationships and features. In addition, using the data collected, researchers are able to use the tool to organise and analyse data using patterns of syntactic relationships and features to either confirm or refute their original hypotheses.

In this paper, we further develop the notion of knowledge-graph-based dynamic survey system that responsively selects questions from a larger pool provided by the researcher. Prioritisation of questions is based on interaction of researchers' hypotheses and participants' input, allowing optimisation of data quality and user responses. Although the proposed survey system is built for linguistic judgements for the sake of evaluation, an architecture of this type could in principle be applied to many other areas of social science research, such as monitoring community opinion on government response to protests, in order to identify finer-grained subgroups of opinions and dependencies amongst opinions. To enable such domain adaptation, per-subject optimisations or randomisation should be in place, in particular when the target issue is sensitive to research bias, such as in the case of assessing response to protests, where findings might influence future policy recommendations. The main contributions from this work include:

- We propose a dynamic survey system that improves on traditional surveys, which typically have a static, fixed-question structure. This system uses knowledge input by the researcher and structured into a knowledge graph to serve appropriate questions to participants rather than presenting all respondents with an identical set of questions.
- We propose two responsive sentence selection algorithms within the dynamic system, which enable hypothesis testing and fine-grained analysis of features and patterns of individual responses. The algorithms are based on a general purpose survey ontology and a domain specific ontology for social science study, such as a Linguistic Feature ontology.
- We present two extensive case studies in Linguistics to evaluate the effectiveness of the proposed algorithms for collecting judgements on a particular grammatical construction. Our evaluation shows that the algorithms have a positive impact on not only the quality of the selected survey questions, but also the number of survey questions needed in the surveys.

The rest of the paper is organised as follows. In Section 2, we review background on knowledge graph and Linguistics, providing examples of the constructions of interest. In Section 3, the requirements of the system are presented. In Section 4, the survey and linguistic feature ontologies are discussed; we detail two alternative algorithms which guide the responsive sentence selection. Several hypotheses are proposed and evaluated with respect to two case studies, as reported in Section 5. Related work is discussed in Section 6. Section 7 concludes the paper.

2. BACKGROUND

2.1 Knowledge Graph

The use of knowledge graph [7, 8] has become popular in knowledge representation and knowledge management applications widely across search [9, 10, 11], recommendation [12], medical informatics [13, 14], finance [15], science [16, 17, 18], media [19], software engineering [20, 21, 22, 23] and industrial domains [7, 24]. In 2012, Google coined the term 'Knowledge Graph (KG)' with a blog post titled "Introducing the knowledge graph: things, not strings". The Knowledge Graph was added to Google's search engine and a 'Knowledge Panel' was added to the search results page. Since then, knowledge graphs have been widely used in the world's leading IT companies.

Formally, a knowledge graph $\mathcal{G} = (\mathcal{D}, \mathcal{S})$ consists of a data sub-graph $\mathcal{D}$ of interconnected typed entities and their attributes as well as a schema sub-graph $\mathcal{S}$ that defines the vocabulary used to annotate entities and their properties in $\mathcal{D}$. Facts in $\mathcal{D}$ are represented as triples of the following two forms:

- *property assertion* (h, r, t), where h is the head entity, and r the property and t the tail entity; e.g., (ACMilan, playInLeague, ItalianLeague) is a property assertion.
- *class assertion* (e, rdf:type, C), where e is an entity, rdf:type is the instance-of relation from the standard W3C RDF specification and C is a class; e.g., (ACMilan, rdf:type, FootballClub) is a class assertion.

A scheme sub-graph $\mathcal{S}$ includes Class Inclusion axioms C $\sqsubseteq$ D, where C and D are class descriptions, such as the following ones: T | $\bot$ | A | $\neg C$ | $C \sqcap D$ | $\exists r.C$ | $\leq nr$ | $= nr$ | $\geq nr$, where T is the top class (representing all entities), $\bot$ is the bottom class (representing an empty set), A is a named class r, r is a property and n is a positive integer. For example, the types of River and City being disjoint can be represented as River $\sqsubseteq\neg$ City, or River $\sqcap$ City $\sqsubseteq\bot$. Schema of knowledge graphs are based on Description Logics [25]. If schema are too expressive, there are well known techniques to simplify schema of knowledge graphs, such as approximations [26, 27] and forgetting [28, 29, 30]. There are uncertainty mechanisms for supporting knowledge graph schema, such as fuzzy extensions [31, 32] and possibilistic extensions [33, 34].

As a means of explicit, formal knowledge representation, a KG facilitates inference, querying, and explanation; a KG adopts the open world assumption and can reason with negation and uncertainty. In addition, KGs are operationalised in machine-readable forms which can be straightforwardly customised, transmitted, reused, and extended [7, 8]. While these are attractive properties, they are not exploited in this paper, though they could be in future research.

We use the KG to represent explicit linguistic knowledge to guide the course of the survey. In effect, specific data from the KG can be used to trigger subsequent explorations in the survey.

2.2 Linguistic Background

Theoretical linguists working on morphosyntax, the structure of words and sentences, may use questionnaires to gather data needed to investigate grammatical structure within a given language or dialect. Surveys of this type seek *grammaticality judgements*, determinations of how *well-formed* sentences are, based on native speakers' knowledge of the language [35]. Data of this type allow linguists to describe and define the parameters of a natural language grammar as it is spoken. As such, native speaker judgements of grammaticality are especially important in the study of "Non-Standard' sentence constructions, which differ from a more widely used" 'Standard' norm, allowing researchers to establish the extent of syntactic variation within a language.

Moreover, although participants' judgements might cluster in particular patterns, there may also be a level of individual variation that is obscured by global measurements of grammaticality. 'Naive' native speaker respondents also often make judgements based on *acceptability*, subject to the influence of factors such as pragmatic plausibility, which is whether a sentence can be used in a particular setting, rather than pure *grammaticality*, which is correctness of (morpho)syntactic structure [36]. Thus, to measure such interspeaker variation, researchers may wish to seek clarification through systematic follow-up questions that are tied to specific grammatical features; these can be difficult to serve in current online data collection approaches.

Our proposed system was evaluated in a use case on the grammaticality of the Alternative Embedded Passive (AEP) [37], which consists of a verb such as *need/want/like* followed directly by a passive participle, in contrast to Standard Embedded Passives (StEP), in which the passive participle is preceded by the

non-finite passive auxiliary *to be*. For example, we might compare speaker grammaticality judgements of the following:

- *The dog needs walked* (AEP);
- *The dog needs* ***to be*** *walked* (StEP).

Linguistic acceptability was tested by giving respondents a binary choice over each question, where 0 stands for *this sentence sounds strange to me* and 1 stands for *this sentence sounds good to me*. In the initial instructions participants were told that a 'strange' sentence would be something they could not say themselves and would be surprised to hear someone else say, while a 'good' sentence would be something they could say themselves or would not be surprised to hear someone else say.

Previous linguistic studies on the AEP (without the presence of *to be*) point to *need* being the most commonly used main verb, followed by *want* and *like* [38]. In addition, inanimate subjects seem to be more acceptable with the use of *want* and *like* in the AEP than the StEP [37]. However, these findings are based on studies conducted only on the North American population using American English, and therefore may not apply to Scottish and Northern Irish speakers who use the AEP. We will outline relevant hypotheses in Section 5.

3. REQUIREMENTS FOR AN INFORMATIVE AND DYNAMIC SYSTEM

In this section, we outline two main requirements for an informative and dynamic system.

An Informative System. Respondents should be asked a sufficient and reasonable number of relevant questions about their grammars and few irrelevant or redundant questions. Sufficient means that we ask enough questions to address our hypotheses, and reasonable means we limit the number of questions we ask each individual respondent to about 30, consistent with other "dialect" surveys [37, 39]. As for relevance, we consider the following aspects:

- Testing the variables by using the researcher's annotations to ask questions including all linguistic features;
- Testing the hierarchy of acceptability of different linguistic features;
- Validation of grammatical points by checking responses about grammar rather than extraneous factors such as pragmatic plausibility; and
- Filtering of questions to avoid those that are known, from prior questions, not to be appropriate to the speaker.

Implementation of these aspects is discussed in Section 4.

A Dynamic System. The dynamic system should serve the purposes of the informative system. In other words, the algorithms which deliver the questions to respondents ought to do so in such a way as to realise the requirements which make the system informative. This design is in contrast to typical grammaticality

judgement questionnaires, where the same questions are asked in every survey, not taking into account the participants' responses. In these traditional static surveys the order of the questions is predefined or entirely randomised in advance, and therefore cannot be changed as the survey is conducted. The fixed presentation of questions does not allow for a more tailored experience for the respondent and does not allow for user feedback in the form of comments to be taken into account in real time. Additionally, the number of questions is limited, which means that a researcher may only be able to cover a select few variables of interest.

4. KNOWLEDGE GRAPH AND ALGORITHMS

In this section, we will present some knowledge-graph-based algorithms for informative and dynamic survey systems. Such survey systems are based on the notion of responsive sentence selection. In other words, the proposed survey system is able to dynamically select the next survey question, depending on the judgement given for the previous question. We will first present the two ontologies as the schema of the knowledge graph of the survey system, which allows the kind of responsive sentence selection to be presented in the algorithms in Section 4.2.

4.1 Knowledge Graph

Two key ontologies are designed for the proposed system: a general purpose Survey Ontology and a domain specific ontology, such as a Linguistic Feature Ontology.

The Survey Ontology contains classes such as SurveyQuestion, AnswerOption, SurveyAnswer and User, Participation, and Hypothesis. It contains properties, such as hasSurveyUser, hasSurveyQuestion and hasSurveyAnswer. We refer the reader to [6] for more details of the Survey Ontology.

The Linguistic Feature Ontology has classes such as Sentence, POS, Subject (Subject $\sqsubseteq$ POS), AnimateSubject (AnimateSubject $\sqsubseteq$ Subject), InanimateSubject (InanimateSubject $\sqsubseteq$ Subject), DefiniteSubject (DefiniteSubject $\sqsubseteq$ Subject), IndefiniteSubject (IndefiniteSubject $\sqsubseteq$ Subject), Verb (Verb $\sqsubseteq$ POS), MainVerb (with instances *need/want/like*, MainVerb $\sqsubseteq$ Verb), AEP (AEP $\sqsubseteq$ POS) and StEP (StEP $\sqsubseteq$ POS). The Linguistic Feature Ontology has properties such as hasPOS and hasString.

When a linguistic researcher annotates survey questions (such as the one containing Sentence S1, *The dog needs walked*), a set of statements will be constructed in the knowledge graph:

- (theDog, rdf:type, DefiniteSubject), (S1, hasPOS, theDog);
- (theDog, rdf:type, AnimateSubject);
- (need, rdf:type, MainVerb), (S1, hasPOS, need); and
- (walked, rdf:type, AEP), (S1, hasPOS, walked).

In the algorithms to be presented in the next section, survey questions sharing the same set of features are classified into the same sentence group. In what follows, we will illustrate how to use the Linguistic Feature Ontology for the classification. For example, given the above statements, we can, e.g., classify the Sentence S1 as an instance of Sentence ⊓∃ hasPOS.DefiniteSubject ⊓∃ hasPOS.AnimateSubject ⊓∃ hasPOS.AEP (which says, S1 is a Sentence that has a DefiniteSubject, an AnimateSubject and contains an AEP). Based on the classification, questions sharing the same set of features can be put into the same sentence group (*cf*. Section 4.2).

In the algorithms to be presented in Section 4.2, the user judgement of the current question is used to pick the next suitable survey question. Thus, it is important that intelligent survey systems can turn user judgements into knowledge graph assertions, so as to activate the responsive sentence selection process. For example, if a User U1 accepts S1, the survey system will have the following extra assertions in the system knowledge graph:

- (P1, rdf:type, Participation);
- (P1, hasSurveyUser, U1), (U1, rdf:type, User);
- (P1, hasSurveyQuestion, S1), (S1, rdf:type, Sentence); and
- (P1, hasSurveyAnswer, *accepted*).

Such information is also useful for the data analysis stage, when the researcher could look into how each individual participant goes through each of their survey questions, including their comments.

4.2 Algorithms

In this sub-section, we will present two algorithms that are able to responsively select a sentence for the next question, with the help of the sentence classification discussed in the previous sub-section (*cf*. the discussion of the survey sentence S1). In other words, subsequent sentences are contingent on the ontological classifications of the prior sentence. In order to illustrate some technical details of our responsive sentence selection algorithms, we choose a linguistic use case for illustration purposes, while we try to keep the algorithms as general as possible. Essentially, while the algorithms will need to be adapted if they are to be reused in another scientific survey system, the adaptations ought to be relatively minimal.

4.2.1 Responsive Sentence Selection

In order to pick the right question(s) for a given participant, an intelligent survey system selects survey questions responsively; in other words, the system tries to pick the next survey question based on the answer for the current one. The main procedure for responsive sentence selection is presented in Algorithm 1.

Before this procedure, there are two pre-processing steps.

Step 1: Grouping. All questions sharing the same set of features are put into the same sentence group. For example, in the linguistic case, the linguistic features include the choice of the main verb, namely *need,*

like, want, as well as whether the subject is *Animate/Inanimate* or *Definite/Indefinite*. Under this setting, one sentence group, e.g., could include sentences only with *need*, and *Animate* and *Definite* subjects. Along with these variables, the presence/absence of the non-finite passive auxiliary *to be* gives a total of 3 * 2 * 2 * 2 = 24 possible combinations of relationships between features. These variables have been explored in previous work on this construction [37]. The linguistic researchers designed six sentences for each of the above 24 combinations, resulting in 24*6=144 sentences, which are grouped into 12 family groups. Each family group has an AEP family of six sentences and a StEP family of six sentences.

Step 2: Ranking. The full list of survey sentences should be ranked based on some pilot study, in terms of some measure for the scientific survey system. Such ranking can be updated later on based on new survey results. In the linguistic case, all the 144 survey questions are ranked in terms of acceptability, as determined by previous survey results. This means that, in each sentence group, sentences are ranked, from more acceptable to less acceptable. We could define the acceptability of sentence groups and family groups as the average acceptability of the sentences in the group. This allows us to rank family groups as well. The intelligent scientific survey system would deal with the most acceptable family group first, then the next acceptable one, and so on, prioritising sentence types that speakers are most likely to accept.

Algorithm 1. Responsive Sentence Selection

Input:
g: current family group, which contains a set of ranked AEP sentences g.aep and a set of ranked StEP sentences g.step;
x: the top percentage of sentences to be considered;
Output:
result: the set of judgement results of selected questions from g

```
result ← nil;
s1 ← random-top(g.aep,x); //randomly select a sentence from the top x% of questions as the first question
result ← result ∪ (s1, judgement(s1));
if judgement(s1) = rejected then
    s2 ← random-top(g.aep,x); //randomly select another sentence from the top x% of questions as the second question
    result ← result ∪ (s2, judgement(s2));
    if judgement(s2) = accepted then
        s3 ← random-bottom(g.aep,x); //randomly select a sentence from the bottom (100 − x)% of questions as the third question
        result ← result ∪ (s3, judgement(s3));
    end
else
    s2 ← random-bottom(g.aep,x); //randomly select a sentence from the bottom (100 − x)% of questions as the second question
    result ← result ∪ (s2, judgement(s2));
end
s ← random-top(g.step,x);//randomly select a sentence from the top x% of questions as the final question from the StEP family
result ← result ∪ (s, judgement(s));
return result;
```

Within groups the prioritisation of sentences with the highest acceptability is intended to minimise the influence of factors other than grammaticality on judgements: e.g., if sentences with the same morphosyntactic features have different levels of pragmatic plausibility we would seek to present the most pragmatically plausible sentences (as determined by the baseline survey) to participants in order to encourage judgements based on the combination of morphosyntactic features under investigation. The prioritisation of more acceptable groups is intended to promote interest in participants who may be discouraged by repeated presentation of sentences they deem "wrong". It is also in line with the hypotheses for the particular grammatical construction under study (see Section 5).

After the two pre-processing steps, the key procedure for responsive sentence selection is presented in Algorithm 1, which takes as inputs the current family group *g*, which includes two sub-groups, the AEP sub-group *g*.aep and the StEP sub-group *g*.step, as well as a ratio *x*, which represents the top percentage of sentences to be considered. The output of Algorithm 1 is the set of judgement results of selected questions from *g*. Such outputs will be stored in the scientific survey system for future processing, such as hypotheses testing by the scientists, as discussed in our previous work on knowledge graph based intelligent survey systems [18].

As mentioned earlier, the presented Algorithm 1 is tailored to the linguistic use case, which has 144 survey sentences, so as to illustrate some of the technical details. In the chosen linguistic use case, the key challenge is determining how to select some of the 144 sentences for a survey, which typically includes about 30 question slots.

The question selection procedure in Algorithm 1 involves three ideas:

- (A1.1) The survey questions are classified into a few categories, for allocating survey questions. In the linguistic use case, the 144 sentences are divided into 12 family groups[①], each of which has an AEP sub-group and a StEP sub-group. Since StEP questions are a Standard form of English and are mainly served as baselines, only one slot (line 15) is available for StEP questions within each family group, while up to 3 slots (lines 4–14) are available for AEP sentences.
- (A1.2) Since there are fewer sentence slots than candidate sentences, we can make use of the ranking of acceptability to further divide these sub-groups into more acceptable parts (top *x*% within sub-group, lines 2, 5 and 15) and less acceptable parts (bottom (100 – *x*)% within sub-group, *cf*. lines 8 and 12). *x* is an input parameter in the algorithm.
- (A1.3) Instead of covering every one of the 144 sentences, Algorithm 1 randomly selects the next sentence based on user judgements of the current sentence (lines 4 and 7), resulting in each speaker being asked to make judgements on 3–4 sentences per family group. That means in the best situation, Algorithm 1 can deal with eight family groups (in the case that three sentences are enough for each family group), while in the worst case, it can deal with six family groups (in the case that each family

① Note that in Algorithm 2, there are six family groups due to a different feature setting.

group needs four sentences). This dynamic approach is significantly better than the static solution where only two family groups can be handled, since each family group has 12 sentences. In other words, Algorithm 1 on average helps reduce (12 – 3.5)/12 * 100% = 70.8% of questions.

4.2.2 Responsive Sentence Selection with Comments

Although Algorithm 1 has a reasonable reduction rate of questions, it cannot guarantee that the survey will deal with all of the family groups. In order to address this issue, we will present Algorithm 2, which can deal with all of the family groups, and more.

Like Algorithm 1, Algorithm 2 presented here is also tailored to a given linguistic use case for illustration purposes, in order to provide some of the technical details on dealing with comments. In this setting, the linguistic researchers decided to simplify and drop the distinction of *Definite/Indefinite* Subjects, resulting in six family groups, each of which has an AEP family of 12 sentences and a StEP family of 12 sentences.

A key notion here is the capability to deal with comments, which are provided by participants in a text box that appears whenever a participant selects a "reject" judgement. There are different potential capabilities for dealing with comments, depending on the requirements for individual surveys. In Algorithm 2, we cover the following capabilities:

- (C1) Detection of the types of comments. By making use of natural language processing techniques, one could come up with different patterns for recognising the types of comments. For example, line 9 uses the comment type StEP (Standard Embedded Passive), which contains the key phrase "to be". If two sentences in an AEP sub-group are rejected (lines 4 and 7), and if the comments of both s1 and s2 are with type StEP, then all the sentences in the StEP sub-group of this family group will be accepted (line 10). More generally, there could be different NLP techniques for type detection.
- (C2) Detection of some keywords in comments. In the linguistic survey, the labels of all instances of MainVerb (*cf*. Section 4.1), i.e., "need", "want" and "like", are key information for both AEP and StEP. For example, in line 12, commentMainV is used to detect main verbs from the comments. If two sentences in an AEP sub-group are rejected (lines 4 and 7), and if the comments of both s1 and s2 are *n*ot with the type StEP and the main verb in the comment is different from the main verb from the current family group (line 12), then the AEP group of sentences using the main verb given in the comment will be accepted.
  If they are not the same, there is a good chance that the participant is proposing to replace the current main verb with the one mentioned in the comment.
- (C3) Retrieval of similar sentences. This is a useful feature to allow some batch operations. For example, in the linguistic survey, if the main verb from the comment is different from the main verb in the current group *g*, then the AEP group of sentences using the main verb mentioned in the comment will be accepted (line 13); in other words, there is no need to ask these questions to the participant, since they have explicitly stated that they can use sentences with this main verb. This operation greatly helps to increase the number of questions that the system can deal with.

Algorithm 2. Responsive Sentence Selection with Comment Understanding

Input:
g: current family group, which contains a set of ranked AEP sentences g.aep and a set of ranked StEP sentences g.step;
x: the top percentage of sentences to be considered in the first attempt; SV: a queue of additional sentences with extra variables to be used;
SR: a queue of additional relaxed sentences to be used;
Output:
result: the set of judgement results of selected questions from g

```
1  result ← nil;
2  s1 ← random-top(g.aep,x); //randomly select a sentence from the top x% of questions as
     the first question
3  result ← result ∪ (s1, judgement(s1));
4  if judgement(s1) = rejected then
5      s2 ← random-top(g.aep,x); result ← result ∪ (s2, judgement(s2));
6      // If both s1 and s2 are rejected for the same reason, the next additional relaxed
         sentence
7      if judgement(s2) = rejected and commenttype(s1) = commenttype(s2) then
8          s3 ← next(SR); result ← result ∪ (s3, judgement(s3));
9          if commenttype(s1)=StEP then
10             foreach sentence si in g.step do result ← result ∪ (si, accepted) ;
11         end
12         if commenttype(s1) = other and commentMain(s1) ≠ g.mainWord then
13             foreach sentence si in similar(g, commentMain(s1)) do
                 result ← result ∪ (si, accepted) ;
14         end
15     else
16         s3 ← random-top(g.aep,x); result ← result ∪ (s3, judgement(s3));
17     end
18 else
19     s2 ← next(SV); s3 ← next(SV); //Present the next 2 sentences with an additional
         variable
20     result ← result ∪ (s2, judgement(s2)) ∪ (s3, judgement(s3));
21 end
22 s ← random-top(g.step,x);//randomly select a sentence from the top x% of questions as
     the final question from the StEP family
23 result ← result ∪ (s, judgement(s));
24 return result;
```

To summarise, Algorithm 2 allows the use of comments (lines 9 and 12) provided by respondents to speed up the decision process: each family has four sentence slots; if some of these slots are not needed, additional sentences can be considered, such as those in *SV* and *SR* (lines 19 and 8, respectively) for each individual participant. Two sentences in *SV* will be asked if the first candidate question s_1 is accepted; these questions have some extra variables, so as to test the limit of acceptability from the participant. On the other hand, if both the first two candidate questions s_1 and s_2 are rejected for the same reason (line 7), then one related sentence from *SR* will be used, so as to see if some relaxation can help to make it more acceptable. Consequently, Algorithm 2 allows the consideration of all combinations of variables covered

by the 144 sentences, as well as some additional linguistic features in *SV* and *SR* covered by an additional 18 sentences (*cf*. Section 5.3 for more details of *SV* and *SR*, including coverage of additional linguistic variables).

Note that some linguistic based optimisations and randomisation are used in the two algorithms. In order to adapt these algorithms to another domain, other per-domain optimisations and randomisation should be applied.

5. CASE STUDIES AND EVALUATIONS

We present two case studies and associated evaluations of the results of the algorithms.

5.1 Hypotheses

In order to address the requirements (informative and dynamic) and evaluate a tool that attempts to address them, two versions of the survey were implemented and conducted. The first version, using Algorithm 1, was found not to deliver sufficiently informative results, leading to the development of a second version, with adjustments employed to create Algorithm 2; while the second version is an improvement over the first, we later discuss further refinements.

As described above, our case study examines the use of Alternative Embedded Passives, in which a verb such as *need*, *want* or *like* is followed directly by a passive participle, without the non-finite auxiliary *to be* found in Standard Embedded Passives.

- *The cat needs fed* (AEP)
- *The cat needs* ***to be*** *fed* (StEP)

The AEP has been claimed to be found among speakers in Scotland and Northern Ireland, but there has been little empirical examination of this feature for these populations. We therefore seek to investigate the following hypotheses:

- **Hypothesis 1**: Speakers who use AEP *like* will also use AEP *want*, and speakers who use AEP *want* will also use AEP *need*.
- **Hypothesis 2**: Of speakers who can use both AEP and StEP *want* and *like*, those who accept inanimate subjects with StEP *want* and *like* will also accept inanimate subjects with AEP *want* and *like*.
- **Hypothesis 3**: Some subset of speakers who reject the AEP altogether or for particular main verbs (*need*, *want*, *like*) will accept this construction with modification by an adverb, e.g., *The books need sorted alphabetically*, rather than *The books need sorted*.

5.2 Case Study 1

5.2.1 Experiment Setup

Based on the results from [6], a pool of 144 sentences was divided into 24 families, paired into 12 groups comprising both AEP and StEP sentences. The sentences in each group shared the same set of linguistic features: main verb (*need*, *want*, *like*), subject (in)animacy, and subject (in)definiteness. For instance, the group for *need*, inanimate subject, and definite subject included the following sentences[2].

- *The trees need pruned*
- *The house needs painted*
- *The windows need cleaned*
- *The plant needs to be watered*
- *The garden needs to be tended*
- *This room needs to be tidied*

The sentences were ranked according to their mean ratings in the baseline results from [6], which had 50 participants over six versions, each consisting of 24 sentences covering all combinations of the main verb, (in)definiteness, (in)animacy, and [±*to be*] variables. In the present case study, these sentences were presented to participants according to Algorithm 1.

The family groups were ordered to present those with main verb *need*, followed by those with main verb *want*, followed by those with main verb *like*. This ordering was in line with expectations regarding probability of acceptance, i.e., based on previous work AEP *need* sentences were deemed most likely to be judged grammatical, followed by *want* and *like* sentences. For each rejected sentence participants were asked "What would you say instead?".

Forty-six participants, who were recruited through word of mouth and social media, completed the survey online. Each answered a minimum of 24 questions; those who chose to continue could answer up to 30 questions. At the end of the survey participants were provided with an individualised map[3] comparing their answers on one of the AEP sentences (without *to be*) with other users who had made judgements on sentences with the same set of linguistic features; in order to facilitate this mapping and provide information on the geographic distribution of linguistic features to researchers, survey-takers were asked to answer a number of optional demographic questions before making grammaticality judgements. This map output to respondents was intended as a reward for participation.

5.2.2 Hypothesis Testing

The survey system has allowed examination of Hypothesis 1 in relation to individual speakers, rather than just over global percentages. Of 46 participants, 42 accepted AEP *need*. Thirty-eight of these participants

[2] While the sentences may vary in singular/plural subject, this is not a relevant experimental variable, but provided only for variety.

[3] https://knowledge-representation.org/j.z.pan/pub/IndividualisedMap.png

accepted AEP *want*. Ten of these 38 participants accepted AEP *like*. There were no participants who accepted AEP *like* but not AEP *want*, or who accepted AEP *want* but not AEP *need*. These results therefore confirm the hypothesis that acceptance of *want* in this construction is a precondition for acceptance of *like*, and acceptance of *need* is a precondition for acceptance of *want*.

The system also allows the testing of Hypothesis 2, which posits that speakers are more likely to accept inanimate subjects with AEP *want* and *like* than StEP *want* and *like*. Hypothesis 2 applies only to the subset of speakers who have already been shown to accept both StEP and AEP *want* or *like*, as rejection of these forms with inanimate subjects for speakers who do not use them at all will be irrelevant.

Thirty-two speakers gave positive judgements for both StEP and AEP *want*; ten speakers gave positive judgements for both StEP and AEP *like*[4]. Of the 32 StEP/AEP *want* participants, 15 were asked to judge sentences with inanimate subjects for *want* in both constructions; of the 10 StEP/AEP *like* participants, three were asked to judge sentences with inanimate subjects for *like* in both constructions.

For *want*, three out of 15 participants accepted an inanimate subject with AEP *want* but not StEP *want*, while one accepted an inanimate subject with StEP *want* but not AEP *want*, contrary to the expected pattern. The rest of the speakers either accepted or rejected all inanimate subjects in both constructions. For *like*, two out of three speakers accepted an inanimate subject with AEP and StEP *like*, and the other speaker rejected both. The hypothesis was therefore confirmed in 17 out of 18 instances (94%), but for less than half of the participants to whom this hypothesis might apply. This version of the system did not allow the testing of Hypothesis 3.

5.2.3 Dynamicity and Informativity

The dynamic approach used in this survey was effective in testing the *need* > *want* > *like* hierarchy of acceptance for the AEP (Hypothesis 1), in that it allowed all speakers to be asked questions for each of these main verbs. That said, the ordering of questions to prioritise *need* over *want* and *want* over *like* means that speakers may have been asked fewer questions overall about main verb *want* and especially *like*, as depending on their answers they may have been questioned about as few as eight family groups (out of 12). We can nevertheless conclude that the survey was sufficiently informative for Hypothesis 1.

At the same time, the algorithm used in this iteration of the survey, along with the limitation on the number of questions, meant that testing of Hypothesis 2 was limited. Several participants who did not use AEP *like* or *want* at all were asked to give judgements on this construction with an inanimate subject, resulting in the collection of data irrelevant to our hypothesis. Only 15 of the 32 participants who accepted AEP and StEP *want* were asked to give judgements on this verb with inanimate subjects, and only three of the ten speakers accepted AEP and StEP *like* were asked to give judgements on this verb with inanimate subjects; some of them also gave judgements on only a single sentence for the StEP or the AEP with an

[4] There was an overlap amongst these two groups of speakers: eight of the 10 speakers who gave positive judgements for StEP and AEP *like* had also accepted StEP and AEP *want*. In our results we treat their judgements for *like* and *want* with inanimate subjects as separate data points.

inanimate subject. The algorithm used in this iteration of the survey therefore failed to collect optimal data for testing Hypothesis 2, and was thus insufficiently informative.

The dynamic aspect of the survey was therefore only partially successful. It allowed relatively strong confirmation of use of the AEP, as many speakers did not accept all sentences for this construction; had they been asked only a single sentence and rejected it the result would have been a false negative for use of the AEP. In other instances, though, the dynamic presentation of questions meant the survey collected superfluous or insufficient data.

It is important to note that the elicitation of superfluous judgements is a feature inherent to static surveys (i.e., those with a fixed set of questions for all respondents), and so in this respect the dynamic survey was still superior, as it eliminated these irrelevant questions for at least some participants. Unintentionally insufficient coverage of variables is a problem more easily avoided in a static survey, although by nature having a fixed set of questions circumscribes how many linguistic features a researcher can include in a questionnaire of this type. Below we will discuss amendments intended to remedy this problem in a second iteration of the dynamic survey.

5.3 Case Study 2

5.3.1 Experiment Setup

The same set of 144 sentences was used, divided into 12 families, paired into six groups, comprising both AEP and StEP sentences. Division of the groups was based on main verb and subject (in)animacy: subject (in)definiteness was not used as a variable, as it was deemed irrelevant to any hypothesis of interest, although variation in this feature for the sentences remained as an artefact of previous work.

A further six sentences were added to the set of possible questions in order to test Hypothesis 3, with two each for each main verb. These are the 'relaxed' (SR) sentences employed in Algorithm 2:

- ***need***
  Those shelves need dusted thoroughly
  The books need sorted alphabetically;

- ***want***
  He wants repaid completely
  Kittens want stroked gently; and

- ***like***
  The doctor likes consulted promptly
  Her gran likes visited frequently.

Another 12 sentences were also added to test the acceptability of a number of additional variables in conjunction with the AEP. These are the "extra variable" (SV) sentences employed in Algorithm 2:

- **negation**
  Those carpets don't need cleaned
  The pancakes don't need flipped again;

- **purpose clause**
  The screws need tightened to hold the shelf up
  These oranges need peeled so they can be eaten;

- **relative clause**
  Those are the shirts that need ironed
  There are problems that need solved;

- **intervening adverb** (between the main verb and participle)
  The answer needs fully explained
  His mum wants quickly phoned;

- ***by*-phrase**
  My car needs checked by a mechanic
  His injury needs treated by a doctor;

- **prepositional phrase**
  I need picked up at the station; and

- **question**
  Does the door need opened?

These additional linguistic features were included to measure a number of other hypotheses examined in previous work; though they are tangential to the main hypotheses considered in this paper, we will address them briefly below in order to further consider the effectiveness of the survey algorithm in maximising informativity.

In this iteration, the system was coded to recognise comments in response to "What would you say instead?", in particular, the use of *to be* or an alternative main verb *need*, *want* or *like*. When these comments were recognised by the system, sentences with the relevant features were marked as accepted, and the presentation of subsequent sentences was adjusted accordingly. The sentences were presented according to Algorithm 2, again using participants' judgements from the baseline survey for ranking of the 144 original sentences.

Fifty-three participants were recruited through paid social media advertising which targeted users in Scotland and Northern Ireland. Each participant gave judgements on up to 24 sentences and, as in Case Study 1, was presented with an individualised map upon completion of the questionnaire and encouraged to share the survey on social media. Of the 53 participants, 13 (25%) did not complete the survey; the fewest number of questions answered by any participant was 17.

5.3.2 Hypothesis Testing

Again, the system allowed testing of Hypothesis 1, that use of AEP *need* is a precondition for use of AEP *want*, and AEP *want* is a precondition for AEP *like*. Forty-eight participants of 53 accepted AEP *need*, and 42 of these accepted AEP *want*. Of these 42, 16 accepted AEP *like*. A single participant appeared to accept AEP *need* and *like*, but not AEP *want*, thereby contradicting the claim that use of AEP *want* is a prerequisite for use of AEP *like*. However, closer inspection reveals that this participant accepted only one of five AEP *like* sentences they were asked (*Those babies like cuddled*). Their acceptance of AEP *like* is therefore marginal at best, and it is possible that this judgement was given in error. It is also notable that all of the AEP *want* examples this participant rejected have non-human subjects (e.g., *The cow wants milked*), which may have contributed to their rejection of these sentences. This anomaly therefore does not represent a serious challenge to Hypothesis 1, given that the results for this hypothesis are otherwise nearly categorical. We therefore conclude that the findings from Algorithm 2 confirm Hypothesis 1.

Of 35 participants who accepted both AEP and StEP *want*, 34 were asked about these with inanimate subjects. Five accepted an inanimate subject with AEP *want*, but not StEP *want*; two rejected an inanimate subject with AEP *want* but accepted one with StEP *want*. Of 15 participants who accepted both AEP and StEP *like*, 13 were asked about these with inanimate subjects[5]. Two accepted an inanimate subject with AEP *like*, but not *StEP like*; one rejected an inanimate subject with AEP *like* but accepted one with StEP *like*. The rest of the speakers either accepted or rejected all inanimate subjects in both constructions for *want* and *like*. These results therefore support Hypothesis 2, that speakers who accept inanimate subjects with StEP *want/like* will also accept inanimate subjects with AEP *want/like*, predicated on the notion that inanimate subjects are generally more acceptable with these verbs in the AEP. In only three of 47 instances, speakers contradicted the expected pattern; 94% of responses were in line with Hypothesis 2. Moreover, this hypothesis was tested for 47 out of 50 relevant instances (94%), compared with only 18 out of 42 (43%) in Case Study 1.

Hypothesis 3 applies to speakers who reject the AEP with one or more main verbs. Relevant to this notion we can divide our respondents into four categories[6]:

- **No AEP** (5/53)
  Participants who rejected all the initial AEP sentences ('Standard' speakers)

- ***Need* AEP** (4/53)
  Participants who accepted the initial AEP sentences with main verb *need*, but rejected them with *want* and *like*

[5] As in Case Study 1, there was significant overlap between speakers who accepted both StEP and AEP *like* and those who accepted both StEP and AEP *want*: 12 speakers fell into both categories. Again, we treat their judgements as discrete data points, even where the same speaker was asked to assess inanimate subjects for both *want* and *like*.

[6] As discussed above, a single participant was inconsistent with Hypothesis 1, and was therefore excluded from testing of Hypothesis 3, as their status with respect to these categories is unclear.

- ***Need* & *Want* AEP** (27/53)
  Participants who accepted the initial AEP sentences with main verbs *need* and *want*, but rejected them with *like*

- **All AEP** (16/53)
  Participants who accepted the initial AEP sentences with all main verbs.

For each of these categories, the following number of participants were asked to judge a sentence with an additional adverb that they would be predicted not to accept based on their previous answers. These predictions were made either because the participant accepted no AEP sentences, or none for the main verb in the additional adverb sentence. For instance, a *Need* AEP speaker might be asked to judge *Those shelves need dusted thoroughly* and *He wants repaid completely*, but only the latter sentence would be directly relevant to Hypothesis 3, as a speaker from this group would already be expected to accept a sentence with AEP *need*, regardless of the addition of an adverb. In some cases, therefore, a participant was asked to judge additional adverb sentences, but is not counted in the numbers below, as these sentences had main verbs already predicted to be accepted by that speaker.

- **No AEP**: 5/5
- ***Need* AEP**: 4/4
- ***Need* & *Want* AEP**: 22/27

Three participants were presented only with additional adverb sentences that they would be predicted to accept already based on the main verb. The median number of additional adverb sentences for each group (regardless of main verb) was as follows:

- **No AEP**: 5
- ***Need* AEP**: 3.5
- ***Need* & *Want* AEP**: 2
- **All AEP**: 0.

Of the speakers in the relevant groups asked to judge additional adverb sentences, six out of 31 accepted sentences that they would otherwise be predicted to reject:

- **No AEP**: 0/5
- ***Need* AEP**: 1/4
- ***Need* & *Want* AEP**: 5/22.

It is notable that only those speakers who already accepted some AEP sentences accepted additional adverb sentences. Although the No AEP group consisted of only five respondents, each was asked (and rejected) a minimum of three additional adverb sentences. We therefore conclude that the addition of an adverb may improve acceptability for the AEP, but will not make AEP sentences acceptable for speakers who do not already use this construction, and thus presumably do not have it as part of their grammars. We can therefore conclude that these data support Hypothesis 3 for some AEP speakers, but reject the hypothesis for the group of No AEP speakers.

Additionally, only two of the five All AEP speakers accepted sentences with additional adverbs. This low acceptance rate by All AEP speakers may be down to random variation, but could also indicate that for some speakers the addition of an adverb makes an AEP sentence ungrammatical. We would need a larger sample size to make this determination.

As well as additional adverb sentences, 39 out of 53 speakers were asked to judge a sentence with an additional linguistic feature, as described above. There were 12 of these additional sentences, half of which were included twice in the survey. This doubling meant that certain sentences were prioritised. In some instances, it also meant a participant was asked to judge the same sentence twice, thereby giving us insight into the consistency of their judgements, although in practice this happened only a few times. Where a speaker gave the same judgement twice for one sentence it was counted only once; in the single instance where a participant gave different judgements for the same sentence, both have been included in the average rating.

Algorithm 2 ensured that all of the participants asked to judge these additional sentences had been determined to be AEP speakers to some extent. Based on these 39 participants, acceptance for these additional variables was summarized in Table 1.

Table 1. Additional variables.

Feature	Total asked	Speakers per group asked	Average rating
Negation	32	All: 14; *need* & *want*: 15; *need*: 3	0.91
Purpose-clause	29	All: 14; *need* & *want*: 13; *need*: 2	0.76
Relative clause	16	All: 12; *need* & *want*: 4; *need*: 0	0.91
Intervening adverb	10	All: 9; *need* & *want*: 1; *need*: 0	0.5
By-phrase	9	All: 9; *need* & *want*: 0; *need*: 0	1
Prepositional phrase	3	All: 0; *need* & *want*: 3; *need*: 0	1
Question	1	All: 0; *need* & *want*: 0; *need*: 1	0

These results indicate that these linguistic features are acceptable with the AEP for most speakers, with the exception of intervening adverbs and questions. However, there was considerable variation with respect to how many speakers were asked for each one: The features that were tested more were those for which we had two sentences (as opposed to just one for, e.g., questions), and/or sentences that were included twice in the survey, making them more likely to come up for any individual respondent. Due to a lack of sufficient data, we therefore cannot draw meaningful conclusions for all of these linguistic features.

5.3.3 Dynamicity and Informativity

Because the sentences were divided into fewer families, and participants were not questioned about lower-ranked sentences, this survey was more effective in testing both hypotheses considered in Case Study 1. In particular, nearly all participants for whom Hypothesis 2 was relevant (i.e., those who accepted *want* and *like* in both AEP and StEP constructions) were asked to judge sentences with inanimate subjects. The number of participants asked with each algorithm is summarised in Table 2.

Table 2. The number of relevant speakers Hypothesis 2 was tested for.

	Algorithm 1	Algorithm 2
want	47% (15/32)	97% (34/35)
like	30% (3/10)	87% (13/15)

The addition of comment understanding also meant that some questions could be eliminated, as participants' acceptance of StEP forms (with *to be*) and alternative main verbs could be confirmed by their responses to "What would you say instead?". As a result, it was possible to include additional sentences, allowing testing of Hypothesis 3, and evaluating more linguistic features for many respondents.

Out of 36 potential participants for whom testing of Hypothesis 3 was relevant, 31 were asked to judge sentences with additional adverbs; Only five out of 16 for whom Hypothesis 3 was not relevant were asked to judge these sentences. The algorithm was therefore largely successful in maximising informativity for this hypothesis, and minimising collection of superfluous data.

As evidenced by the results given in Table 1, informativity varied for other linguistic features: Negation and purpose-clauses had robust sample sizes, but only three speakers were asked to judge sentences with prepositional phrases, and only one speaker was asked to judge a sentence with a question. This imbalance was partly due to the prioritisation of certain linguistic features through the inclusion of multiple sentences for those features, or having two instances of particular sentences.

While this iteration of the dynamic survey therefore addressed the problem of insufficient coverage for some variables, thereby increasing informativity, this success was not consistent for all features. Presentation of superfluous questions to some participants still remained a problem. In particular, consistently "Standard" speakers, i.e., those who do not use the AEP at all, were repeatedly presented with AEP sentences because their input of *to be* forms for "What would you say instead?" meant that these StEP forms were marked as grammatical, triggering the algorithm to bypass them. These speakers were therefore never presented with StEP sentences, but instead asked more AEP sentences.

This issue was highlighted by a response left on social media by a participant that the questionnaire became "boring" because all of the questions seemed to require "the same grammatical addition" (presumably insertion of *to be*). Decreasing participant interest as the survey progresses therefore remains a problem, although it is notable that some "Standard" speakers did complete the survey; this problem is potentially remedied by reducing the overall number of questions for participants whose answers indicate that they do not use the Non-Standard form, or introducing new variables and/or constructions for such users in order to increase participant engagement and level of informativity for researchers.

5.4 A Note on Comment Understanding

The comment understanding feature of Algorithm 2 was successful at identifying participants' acceptance of particular constructions, thereby allowing us to eliminate questions on these constructions. However, because the system was designed to respond to only certain expected feedback, it failed to recognise the influence of other variables on speaker judgements. Here we give two examples.

One respondent gave ***That*** *tree needs to be pruned* as an alternative to ***The*** *tree needs to be pruned*, and ***Them*** *students want to be taught* as an alternative to ***Those*** *students want to be taught*. These judgements were recorded as rejections of the *to be* form even though the basis on which the speaker rejected the sentences was the form of the determiner (article) used in the subject. In this instance, then, the system failed to recognise a false negative.

Another respondent gave ***At*** *cat* ***needs till*** *be fed* as an alternative to *The cat likes fed*. Here the system recognised the alternative main verb (*needs* instead of *likes*). It again did not recognise the alternative determiner (**at**, a Non-Standard form of *that*). Moreover, it did not recognise *till* as a Non-Standard form of *to*, and therefore did not adapt the questionnaire based on what was in essence a StEP (*to be*) construction.

To summarise, some responses of this type are relevant to the researcher's key hypotheses (such as those from the second respondent above), while others may not be (such as those from the first respondent above). In the future, we plan to develop further to allow for recognition of variables beyond those expected by the researcher in order to further improve the dynamic delivery of the survey and accuracy of results. In general, this is a challenging problem, as it is hard to predict all kinds of feedback and to adjust the survey accordingly. One of the key issues is the lack of existing survey data to train a prediction model.

6. RELATED WORK

6.1 Intelligent Surveys

An intelligent survey system that has already been implemented elsewhere is the *Dynamic Intelligent Survey Engine* DISE [40], which aims to have an approach that is as flexible as possible to creating a survey, while avoiding being restricted. Similarly to our previous system, it uses a wide variety of data methods and an advanced data collection approach with the intent to measure consumer preferences. However, in contrast to our current system, which uses a drag and drop interface for creating surveys, survey creation in their system is done by XML markup language, which may have a rather steep learning curve and thus be cumbersome to use. Furthermore, the system does not allow for conditional triggers to enable a better user experience, nor does it use its knowledge to prioritise the most significant questions first.

6.2 Psycholinguistic Surveys

MiniJudge [41] is a tool specifically designed for linguists working in theoretical syntax to help them design, run and analyse judgement experiments in the minimum amount of time with maximised efficiency and without any prior training. This aim is achieved by "minimalist" experimental syntax, where experiments are conducted on a small participant group, and sets of questions are limited, resulting in quick surveys. It offers automation of statistical analysis of data, and thus is beginner friendly.

WebExp [42] is a software package to run psychological experiments over the Internet and measure the respondents reaction time (latency). The system shows a nuanced approach to collecting latency measurements and replicating lab-based conditions accurately across multiple platforms. Similarly,

PsyToolkit[7] has been specifically designed to set up, run and analyse questionnaires and reaction time experiments. Furthermore, the system links the experiments online, so that they can easily be embedded in social media networks and used for participant recruitment [43].

Other psycholinguistic tools include IBEX [44] *Internet Based EXperiments* which focuses on grammaticality judgements. The questionnaire presents the sentences in a variety of ways; the *FlashSentence* method where sentences are "flashed" to the participant for only a limited amount of time, and the *DashedSentence* method where the sentences are presented either chunk-by-chunk or word-by-word. Another system, *Wordlikeness*, allows researchers to design questionnaires with text, audio, images and video files in order to make the research feasible across different groups with individual languages [45].

Many other popular survey systems are also used for information gathering in the field of linguistics. One of these is *SurveyMonkey*[8], which allows researchers to develop surveys online, deploying them to the community and then analysing collected data. Another one is *Amazon's Mechanical Turk (MTurk)*[9], where participants complete surveys for money. Johnson suggests that using such a platform can provide a large participation-pool with necessary tools to build an experiment in quick and efficient manner [46]. *Turkolizer* [47] and *Turktools* [48] are two tools that run on this crowd sourcing platform. While this approach may potentially present benefits in large-scale experiments, the platform presents only a basic statistical analysis of the data. To do any form of knowledge powered services, for instance syntactic and semantic evaluation of the results, a knowledge structure would have to be implicitly hard-coded. As a result of limitation, the experimental data are difficult to be transmitted, linked or reused.

6.3 Commercial Surveys

Whilst not as related to other intelligent/linguistic surveys, commercial surveys can provide a great insight on other aspects of surveys such as: *user interfaces, security* and *distribution* [49, 50]. These aspects are paramount as their primary goal is to attract as many customers as possible.

6.4 Adaptive Questionnaires

Adaptive questionnaires are widely used to identify students' learning styles. While questionnaires might be effective, they have disadvantages: (1) filling out a questionnaire is time-consuming since questionnaires usually contain numerous questions; (2) learners may lack time and motivation to fill in long questionnaires; and (3) a specialist needs to analyse the answers [51].

Several questionnaire systems have been proposed to mitigate the above stated issues and automatically minimise the number of questions using various algorithms. *AH questionnaire* [52] used decision trees as the main algorithm and managed to reduce the number of questions by over 50% and achieved over 95% accuracy when predicting the students' learning preference. A tool proposed by Nokelainen et al. [53] uses

[7] https://www.psytoolkit.org/
[8] https://www.surveymonkey.com/
[9] https://www.mturk.com/

Bayesian modelling as well as abductive reasoning and accomplished similar question reduction of 50% as in the previous system. More relevant work to our system has been done with *Q-SELECT* [54] using neural network and decision trees to decrease the number of questions by trying to find the least influential question in the survey. Furthermore, it is capable of reordering the questions, and thus provides a personalised questionnaire to the end-user. Recently, their system *T-PREDICT* has been further improved from 35% reduction of questions to over 85% reduction [55]. However, although they cut survey length, none of these approaches addresses how to responsively select the next question based on judgement of the current one, or in a way that is sensitive to any kind of initial speaker classification.

The current survey also differs from these other systems in seeking not an overall reduction of questions for a fixed selection of variables, but an increase in variable coverage within a fixed number of questions. With a maximum of 30 questions, Algorithm 1 covered up to eight family groups (representing combinations of variables) out of 12, in comparison to only two for a non-dynamic version of the survey. Algorithm 2 covered all family groups, as well as at least one additional variable for 74% of respondents (see discussion in Section 5.3). It is notable that this coverage was achieved despite some participants' discontinuation of the survey before answering the maximum number of questions.

7. DISCUSSIONS AND CONCLUSIONS

In this paper, we have presented an approach that introduces benefits of face-to-face surveys or interviews into online survey systems for social science surveys, powered by knowledge understood by both humans and programs. Many social science projects are switching to online survey systems, not only because they are more scalable in terms of participant recruitment, but also because they provide opportunities to apply additional techniques to provide data about other parameters. For example, as many people turn to electronic book reading [56], online survey systems become a natural choice for social scientists who study reading behaviour, e.g., studying comprehension questions by eye-tracking [57], automated detection of regression, forward leaping, and mind wandering.

Our paper addresses a key problem with online survey systems: the lack of flexibility that is otherwise offered by face-to-face surveys. Underpinning the approach is a novel integration of a decision tree with articulated domain knowledge that is structured into a knowledge graph, providing a high degree of flexibility, fine-grainedness, complexity, and extensibility. Using the knowledge graph, we proposed a dynamic approach to the questionnaire component of the survey, yielding more informative results. The questions are ordered by a model based on their importance and relevance to particular hypotheses; in this way, the approach facilitates hypothesis testing and fine-grained analysis of features and patterns of individual responses. Once the questions are ordered, a set of conditional triggers are set to provide a more dynamic experience, which benefits the researcher in maximising the quality and quantity of data collected, and the user in creating a more varied survey. Follow-up questions are asked according to the user accepting or rejecting certain questions.

In the evaluation, we have shown that the dynamic component can have a positive impact on the quality of the data as well as limiting the number of questions asked in the survey. The previous system performed

six different surveys, each of which had 24 questions; a total of 50 people participated in those surveys. Our system achieves comparable results with a similar number of participants for each iteration (46 and 53, respectively), each asked 24–30 questions. In Case Study 2, especially, it improves upon the previous system by increasing the number of questions asked for particular variables and/or increasing the number of variables covered in the same number of sentences. Such improvement is based on the semantic understanding of survey questions enabled by knowledge graphs.

A primary advantage of the proposed approach is that it enables grouping of participants according to their responses, so that participants are not only served appropriate follow-up questions, but their responses to these questions may be analysed in the context of some initial categorisation. In this study, speakers were broadly grouped according to whether they used a specific syntactic construction (the AEP), with assessment of additional variables made for speakers within these groups. In other types of social science surveys, it is easy to imagine that researchers might choose to make initial groupings based on characteristics such as gender, race, socioeconomic status and political affiliation; for instance, in the example of a survey on government response to protests, researchers might deliver different questions depending on how a participant reports having voted in the last election.

On a more nuanced level, the use of a knowledge graph opens the way to further advances in survey delivery and analysis. For instance, it may allow organisation and delivery of questions according to underlying principles not available with adaptive questionnaires using straightforward decision trees. Questions in a social science survey might be tagged thematically to allow follow-up on a particular topic, or to tease out the reasoning behind a response. In the protest example, for instance, a participant might object to a heavy-handed government response because they distrust the motives of particular political figures, because they strongly support the cause of the protesters, or because they ascribe high value to public protest regardless of the cause.

It is also known that survey participants may give different responses according to the way questions are phrased, such as using the words “not allowed” instead of “forbid” [58]. In the linguistic survey the object of investigation was language itself, but tagging of linguistic features, such as specific words or syntactic structures, could equally be applied to questions in other surveys. In this way, a knowledge graph might underpin social science surveys not just in terms of question content and theme, but also form, measuring variation in response to the language of questions, and potentially asking the “same” question in different ways to different participants. Here, too, comment understanding could be employed to redirect the line of questioning if participants input particular key phrases when asked for further comment.

The organisation of researcher/participant input by knowledge graph might also play a role not just in the deployment, but development of questionnaires. Specifically, it is possible that the system could generate further questions (e.g., sentences for judgement) based on the existing KG. We leave this avenue to future work.

The survey approach proposed here thus allows online questionnaires to approach face-to-face interaction in their level of informativity and responsiveness, as well as duplicating certain advantages of interview-based data collection. Deployment of high-quality surveys over the Internet has become especially relevant

in light of the global COVID-19 pandemic, which severely restricts researchers' freedom of movement and in-person contact with participants.

ACKNOWLEDGEMENTS

We are grateful to all of the participants who gave their time to fill out the linguistic questionnaires discussed in this work.

AUTHOR CONTRIBUTIONS

J.Z. Pan (j.z.pan@ed.ac.uk) was the leading contributor for the computer science aspect of the work. He led the writing of the paper and is the primary author of Section 4. E. Edelstein (elspeth.edelstein@abdn.ac.uk) was the leading contributor for the linguistic aspects of the work. She contributed to writing across the sections of the paper and is the primary author of Section 5. P. Bansky (p.pato.otap@gmail.com) was the primary contributor of the intelligent survey system implementation as well as contributing to writing across the sections. With a background in linguistics and computer science, A. Wyner (a.z.wyner@swansea.ac.uk) contributed to the interdisciplinary discussions and developments intrinsic to the paper as well as to writing across the sections.

REFERENCES

[1] Van Selm, M., Jankowski, N.W.: Conducting online surveys. Quality & Quantity 40, 435–456 (2006)
[2] Stoop, I., Harrison, E.: Classification of surveys. In: Gideon, L. (ed.) Handbook of Survey Methodology for the Social Sciences, pp. 7–21. Springer, Berlin (2012)
[3] Stoop, I., Harrison, E.: Repeated cross-sectional surveys using FTF. In: Gideon, L. (ed.) Handbook of Survey Methodology for the Social Sciences, pp. 249–276. Springer, Berlin (2012)
[4] Lewis-Beck, M., Bryman, A., Liao, T.F.: The Sage Encyclopedia of Social Science Research Methods (2004). Available at: https://www.researchgate.net/publication/241831065_The_Sage_Encyclopaedia_Social_Science_Research_Methods. Accessed 3 July 2021
[5] Trochim, W.: Research methods knowledge base (2007). Available at: https://www.researchgate.net/publication/243783609_The_Research_Methods_Knowledge_Base. Accessed 3 July 2021
[6] Soares, R., et al.: Knowledge driven intelligent survey systems for linguists. In: Joint International Semantic Technology Conference, pp. 3–18 (2018)
[7] Pan, J., et al. (eds.): Exploiting linked data and knowledge graphs in large organisations. Springer, Berlin (2016)
[8] Pan, J., et al.: Reasoning Web: Logical foundation of knowledge graph construction and querying answering. Springer, Berlin (2017)
[9] Pan, J.Z., Taylor, S., Thomas, E.: Reducing ambiguity in tagging systems with folksonomy search expansion. In: The Proceedings of the 6th European Semantic Web Conference (ESWC2009), pp. 1–15 (2009)
[10] Dietz, L., Kotov, A., Meij, E.: Utilizing knowledge graphs for text-centric information retrieval. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), pp. 1387–1390 (2018)
[11] Gu, Y., et al.: Relevance search over schema-rich knowledge graphs. In: Proceedings of the 12th ACM International WSDM Conference (WSDM2019), pp. 114–122 (2019)

[12] Wang, X., et al.: KGAT: Knowledge graph attention network for recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2019), pp. 950–958 (2019)
[13] Wu, H., et al.: Knowledge driven phenotyping. In: Proceedings of Medical Informatics Europe (MIE 2020), pp. 1327–1328 (2020)
[14] Tripodi, I.J., et al.: Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicology in Vitro 66, 104877 (2020)
[15] Deng, S., et al.: Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In: Proceedings of the World Wide Web Conference (WWW 2019), pp. 678–685 (2019)
[16] Xu, H., Giunchiglia, F.: Sko types: An entity-based scientific knowledge objects metadata schema. Journal of Knowledge Management 19(1), 60–70 (2015)
[17] Auer, S., et al.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (WIMS 2018), pp. 1327–1328 (2018)
[18] Edelstein, E., et al.: Knowledge-driven intelligent survey systems towards open science. New Generation Computing 38, 397–421 (2020)
[19] Pan, J.Z., et al.: Content based fake news detection using knowledge graphs. In: Proceedings of the International Semantic Web Conference (ISWC2018), pp. 669–683 (2018)
[20] Pan, J.Z., et al. (eds.): Ontology-driven software development. Springer, Berlin (2013)
[21] Pan, J.Z., Zhao, Y. (eds.): Semantic Web enabled software engineering. IOS Press, Amsterdam (2014)
[22] Siegemund, K., et al.: Towards ontology-driven requirements engineering. In: Proceedings of the International Workshop on Semantic Web Enabled Software Engineering (SWESE2011), pp. 1–15 (2011)
[23] Taylor, S., et al.: Reasoning driven configuration of linked data content management systems. In: Proceedings of the 3rd Joint International Conference on Semantic Technologies (JIST 2013), pp. 429–444 (2013)
[24] Bader, S.R., et al.: A knowledge graph for industry 4.0. In: Proceedings of the 17th Extended Semantic Web Conference (ESWC 2020), pp. 465–480 (2020)
[25] Baader, F., et al. (eds.): The description logic handbook: Theory, implementation, and applications. Cambridge University Press, Cambridge (2003)
[26] Pan, J.Z., Thomas, E.: Approximating OWL-DL ontologies. In: The Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07), pp. 1434–1439 (2007)
[27] Pan, J.Z., Ren, Y., Zhao, Y.: Tractable approximate deduction for OWL. Artificial Intelligence 235, 95–155 (2016)
[28] Wang, Z., et al.: Forgetting for knowledge bases in DL-Lite. Journal of Annals of Mathematics and Artificial Intelligence 58(1-2), 117–151 (2010)
[29] Lutz, C., Wolter, F.: Foundations for uniform interpolation and forgetting in expressive description logics. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 117–151 (2011)
[30] Wang, K., et al.: Eliminating concepts and roles from ontologies in expressive descriptive logics. Computing Intelligence 30(2), 205–232 (2014)
[31] Pan, J.Z., et al.: f-SWRL: A fuzzy extension of SWRL. Journal of Data Semantic 6, 28–46 (2006)
[32] Stoilos, G., Stamou, G.B., Pan, J.Z.: Handling imprecise knowledge with fuzzy description logic. In: The Proceedings of the 2006 International Workshop on Description Logics (DL2006), pp. 119–126 (2006)
[33] Qi, G., Pan, J.Z., Ji, Q.: A possibilistic extension of description logics. In: Proceedings of 2007 International Workshop on Description Logics (DL2007), pp. 1–8 (2007)
[34] Qi, G., et al.: Extending description logics with uncertainty reasoning in possibilistic logic. International Journal of Intelligent Systems 26(4), 353–381 (2011)
[35] Schütze, C.T.: The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Language Science Press, Berlin (2016)

[36] Leivada, E., Westergaard, M.: Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. Frontiers in Psychology 11, 364 (2020)
[37] Edelstein, E.: This syntax needs studied. In: Micro-syntactic variation in North American English, pp. 242–268 (2014)
[38] Murray, T.E., Simon, B.L.: At the intersection of regional and social dialects: The case of like + past participle in American English. American Speech 77(1), 32–69 (2002)
[39] Katz, J.: The British-Irish dialect quiz. New York Times, 15 February 2019.
[40] Schlereth, C., Skiera, B.: Dise: Dynamic intelligent survey engine. In: Diamantopoulos, A., Fritz, W., Hildebrandt, L. (eds.) Quantitative Marketing and Marketing Management, pp. 225–243. Springer, Berlin (2012)
[41] Myers, J.: Minijudge: Software for small-scale experimental syntax. International Journal of Computational Linguistics & Chinese Language Processing 12(2), 175–194 (2007)
[42] Keller, F., et al.: Timing accuracy of web experiments: A case study using the webexp software package. Behavior Research Methods 41(1), 1–12 (2009)
[43] Stoet, G.: Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. Teaching of Psychology 44(1), 24–31 (2017)
[44] Drummond, A.: ibex 0.3.7 manual (2013). Available at: https://github.com/addrummond/ibex/blob/master/docs/manual.md. Accessed 3 July 2021
[45] Chen, T.Y., Myers, J.: Worldlikeness: A web-based tool for typological psycholinguistic research. University of Pennsylvania Working Papers in Linguistics 23(1), 4 (2017)
[46] Johnson, D.R., Borden, L.A.: Participants at your fingertips: Using Amazon's Mechanical Turk to increase student-faculty collaborative research. Teaching of Psychology 39(4), 245–251 (2012)
[47] Gibson, E., Piantadosi, S., Fedorenko, K.: Using Mechanical Turk to obtain and analyze English acceptability judgments. Language and Linguistics Compass 5(8), 509–524 (2011)
[48] Erlewine, M.Y., Kotek, H.: A streamlined approach to online linguistic surveys. Natural Language & Linguistic Theory 34(2), 481–495 (2016)
[49] Capterra: Survey software buyers' guide. Available at: https://www.capterra.com/survey-software/#buyers-guide, 2019. Accessed 5 Mar 2019
[50] Software advice, buyer's guide. Available at: https://www.softwareadvice.com/za/survey/#buyers-guide. Accessed 23 April 2019
[51] Abernethy, J., Evgeniou, T., Vert, J.-P.: An optimization framework for adaptive questionnaire design (2004). Available at: https://sites.insead.edu/facultyresearch/research/doc.cfm?did=1440. Accessed 3 July 2021
[52] Ortigosa, A., Paredes, P., Rodriguez, P.: Ah-questionnaire: An adaptive hierarchical questionnaire for learning styles. Computers & Education 54(4), 999–1005 (2010)
[53] Nokelainen, P., et al.: Implementation of an adaptive questionnaire. In: Proceedings of the ED-MEDIA Conference, pp. 1412–1413 (2001)
[54] Mwamikazi, E., et al.: An adaptive questionnaire for automatic identification of learning styles. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 399–409 (2014)
[55] Mwamikazi, E., et al.: A dynamic questionnaire to further reduce questions in learning style assessment. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 224–235 (2014)
[56] Nicholas, D., et al.: UK scholarly e-book usage: A landmark survey. Aslib Journal of Information Management 60(4), 311–334 (2008)
[57] Beymer, D., Orton, P.Z., Russell, D.M.: An eye tracking study of how pictures influence online reading. In: Proceedings of IFIP Conference on Human-Computer Interaction, pp. 456–460 (2007)
[58] Converse, J.M., Presser, S.: Survey questions: Handcrafting the standardized questionnaire. Sage Publications, Thousand Oaks (1986)

AUTHOR BIOGRAPHY

Jeff Z. Pan has a PhD in Computer Science (University of Manchester, 2004). He is a Reader in the School of Informatics at the University of Edinburgh and a Chair of the Knowledge Graphs Group of the Alan Turing Institute in the UK. His research focuses primarily on knowledge representation and artificial intelligence, in particular on knowledge graph based learning and reasoning, and knowledge based natural language understanding and generations, as well as their applications. As the Chief Editor of the first two books on knowledge graphs, he is an Associate Editor of the *Journal of Web Semantics* (JoWS) and a Programme Chair of the International Semantic Web Conference (ISWC 2020), the premier international forum for the Semantic Web and Knowledge Graph communities. For more information about him, see https://knowledge-representation.org/j.z.pan/.
ORCID: 0000-0002-9779-2088

Elspeth Edelstein has a PhD in Linguistics (University of Edinburgh, 2012). She is currently a Senior Lecturer in Language & Linguistics at the University of Aberdeen. Her research focuses on natural language syntax, looking particularly at (morpho)syntactic variation and non-standard syntactic constructions. Notable recent publications include the monograph *English Syntax: A Minimalist Account of Structure and Variation* (Edinburgh University Press, 2020).
ORCID: 0000-0002-1995-2787

Patrik Bansky is currently a Research Software Engineer at Huawei Technologies Research and Development, working on knowledge graph related topics. He completed his Computer Science degree from The University of Aberdeen in 2019.

Adam Wyner has PhDs in Linguistics (Cornell University, 1994) and Computer Science (King's College London, 2008). Currently an Associate Professor at Swansea University and holding a joint position in the School of Law and Department of Computer Science, he lectures and conducts research on artificial intelligence and law. He has published on natural language processing (rule-based and machine learning), information extraction, ontologies, argumentation, controlled languages, case based reasoning, policy consultations, semantic Web, and a machine-readable standard language for legal rules.
ORCID: 0000-0002-2958-3428